Course Project Description **10/17/19**

It's time to start working on your course projects. You may choose from one of two topics (topic 1, 2) below, and communicate your choice to me. In your email (subject line should be "CS 598 Project"), indicate what your team composition is, and which topic you will work on. Please send this email by end of Thursday 10/24.

A written report of the project and a presentation to the instructor will be expected in the finals week. (Specific dates will be finalized later.)

Topic 1: Predicting transcription factor-DNA binding from sequence

Here, you will be given training data that include a set of ChIP peaks, i.e., bound regions of DNA for a particular TF and a set of non-bound regions, and have to build a predictor/classifier that can tell if a sequence (that it hasn't seen before) is bound or not. The "test data" (not to be used in training your classifier) will also be provided. Training + test data for several TFs will be made available. You will obtain these data from the course web page. (It's not there yet, but will, soon.)

Additional reading on this topic may include papers written by participants of the DREAM challenge that was held on the topic two years ago:
https://genome.cshlp.org/content/29/2/281.full
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6327544/

Also, we have covered the CENTIPEDE paper (Pique-Regi et al.) in class, and this is related to though not the exact same problem as the project.

Your project will:
(1) Obtain data for training and testing.
(2) Implement a method for the supervised learning task of predicting TF-DNA binding.
(3) Perform rigorous evaluations of your method, e.g., through cross-validation.

Topic 2: Transcriptional Regulatory Network reconstruction

Here, you will work with the data sets from the Siahpirani & Roy paper (covered in class) to infer TRNs from expression data. They have made their data available at:
https://github.com/Roy-lab/merlin-p_inferred_networks/tree/master/yeast_networks

The above data set includes:
(1) Expression data (folder expression/). You may use all or any subset of these data to reconstruct TRNs. Your choice should be clearly justified/motivated.
(2) Benchmark TRN (folder gold/). You should use each of the three gold-standard TRNs provided here for evaluation purposes. Measures of evaluation may be inspired by the Siahpirani & Roy paper, but you are not limited to their methodology for evaluation.
(3) Prior networks (folder priors/). You may use any or all of these three networks to aid your TRN reconstruction, but it is optional.

The goal of your project is to develop and understand the strengths and weaknesses of a method and its variants. The accuracy of your TRN predictions (by whatever measures you

choose) is only part of the criteria by which your project will be evaluated. Rigor in evaluation and the interestingness of your method and evaluation are also going to be crucial.

Your project will:

(1) Obtain data for analysis and evaluation.
(2) Implement a method for the TRN inference problem.
(3) Perform rigorous evaluations of your method, as well as variants thereof.

The most important references for this project are the two papers we have covered in class:

A prior-based integrative framework for functional transcriptional regulatory network inference. Siahpirani & Roy, NAR 2017.

Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. Vân Anh Huynh-Thu, Alexandre Irrthum, Louis Wehenkel, Pierre Geurts. PLoS One 2010.

Topic 3 (research project)

I am looking for a single team to work on a relatively narrowly-scoped but not very well-defined project on mining the human microbiome for a certain class of compounds. This will be a collaboration with Prof. Satish Nair's lab (Biochemistry & Biophysics), and will require the use of HMMs for scanning genomes/proteomes. I will provide guidance on the methodology here, since there are no great reference points to build the project upon. Strong programming background is necessary. The effort required will be comparable to or a little less than the other two topics.

If you are interested, let me know in your email that this is your top preference and then indicate which topic (1 or 2) you'd like to work on if you are not assigned topic 3. If there are multiple teams that volunteer for topic 3, I will select one of them at my discretion.